ELSEVIER

Journal of Clinical Epidemiology 66 (2013) 398−407

Journal of Clinical Epidemiology

ORIGINAL ARTICLES

Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes

Peter C. Austin[a,b,c,*], Jack V. Tu[a,b,d], Jennifer E. Ho[e,f,g], Daniel Levy[e,f,h], Douglas S. Lee[a,b,i]

[a]*Institute for Clinical Evaluative Sciences, G105, 2075 Bayview Ave, Toronto, Ontario, Canada*
[b]*Institute of Health Management, Policy and Evaluation, University of Toronto, Suite 425, 155 College St., Toronto, Ontario, Canada*
[c]*Dalla Lana School of Public Health, University of Toronto, 6th floor, 155 College St., Toronto, Ontario, Canada*
[d]*Division of Cardiology, Sunnybrook Schulich Heart Centre and Faculty of Medicine, University of Toronto, 2075 Bayview Ave., Toronto, Ontario, Canada*
[e]*National Heart, Lung, and Blood Institute's Framingham Heart Study, 73 Mt. Wayte Ave., Framingham, MA, USA*
[f]*Center for Population Studies, National Heart, Lung, and Blood Institute, 31 Center Dr., Bethesda, MD 20892, USA*
[g]*Department of Medicine, Section of Cardiovascular Medicine, Boston University, 88 East Newton St., C-818, Boston, MA 02118, USA*
[h]*Department of Medicine, School of Medicine, Boston University, 72 E. Concord St., Boston, MA 02118, USA*
[i]*Department of Medicine, University Health Network and Faculty of Medicine, University of Toronto, Room 4NU-482, 200 Elizabeth St., Toronto, Ontario, Canada*

Accepted 25 November 2012; Published online 4 February 2013

Abstract

Objective: Physicians classify patients into those with or without a specific disease. Furthermore, there is often interest in classifying patients according to disease etiology or subtype. Classification trees are frequently used to classify patients according to the presence or absence of a disease. However, classification trees can suffer from limited accuracy. In the data-mining and machine-learning literature, alternate classification schemes have been developed. These include bootstrap aggregation (bagging), boosting, random forests, and support vector machines.

Study Design and Setting: We compared the performance of these classification methods with that of conventional classification trees to classify patients with heart failure (HF) according to the following subtypes: HF with preserved ejection fraction (HFPEF) and HF with reduced ejection fraction. We also compared the ability of these methods to predict the probability of the presence of HFPEF with that of conventional logistic regression.

Results: We found that modern, flexible tree-based methods from the data-mining literature offer substantial improvement in prediction and classification of HF subtype compared with conventional classification and regression trees. However, conventional logistic regression had superior performance for predicting the probability of the presence of HFPEF compared with the methods proposed in the data-mining literature.

Conclusion: The use of tree-based methods offers superior performance over conventional classification and regression trees for predicting and classifying HF subtypes in a population-based sample of patients from Ontario, Canada. However, these methods do not offer substantial improvements over logistic regression for predicting the presence of HFPEF.

Keywords: Boosting; Classification trees; Bagging; Random forests; Classification; Regression trees; Support vector machines; Regression methods; Prediction; Heart failure

Conflict of interest statement: The authors declare that there is no conflict of interest.

Funding: This study was supported by the Institute for Clinical Evaluative Sciences (ICES), which is funded by an annual grant from the Ontario Ministry of Health and Long-Term Care (MOHLTC). The opinions, results, and conclusions reported in this article are those of the authors and are independent from the funding sources. No endorsement by ICES or the Ontario MOHLTC is intended or should be inferred. This research was supported by an operating grant from the Canadian Institutes of Health Research (CIHR) (MOP 86508). Dr Austin is supported in part by a Career Investigator award from the Heart and Stroke Foundation. Dr Tu is supported by a Canada Research Chair in Health Services Research and a Career Investigator Award from the Heart and Stroke Foundation. Dr Lee is a clinician−scientist of the CIHR. The data used in this study were obtained from the Enhanced Feedback for Effective Cardiac Treatment (EFFECT) study. The EFFECT study was funded by a CIHR Team Grant in Cardiovascular Outcomes Research.

* Corresponding author. Institute for Clinical Evaluative Sciences, G1 06, 2075 Bayview Avenue, Toronto, Ontario M4N 3M5, Canada. Tel.: +1 416 480 6131; fax: +1 416 480 6048.

E-mail address: peter.austin@ices.on.ca (P.C. Austin).

0895-4356/$ - see front matter
http://dx.doi.org/10.1016/j.jclinepi.2012.11.008

What is new?

Key findings

- Modern data-mining and machine-learning methods offer advantages for predicting and classifying heart failure (HF) patients according to disease subtype: HF with preserved ejection fraction (HFPEF) and HF with reduced ejection fraction compared with conventional regression and classification trees.
- Conventional logistic regression performed at least as well as modern methods from the data-mining and machine-learning literature for predicting the probability of the presence of HFPEF in patients with HF.

What this adds to what was known?

- Boosted trees, bagged trees, and random forests do not offer an advantage over conventional logistic regression for predicting the probability of disease subtype in patients with HF.

What is the implication and what should change now?

- Conventional logistic regression should remain a standard tool in the analyst's toolbox when predicting disease subtype in patients with HF.
- Analysts interested in classifying HF patients according to disease subtype should use ensemble-based methods rather than conventional classification trees.

1. Introduction

There is an increasing interest in using classification methods in clinical research. Classification methods allow one to assign subjects to one of a mutually exclusive set of states. Accurate classification of disease states (disease present/absent) or of disease etiology or subtype allows subsequent investigations, treatments, and interventions to be delivered in an efficient and targeted manner. Similarly, accurate classification of disease states permits more accurate assessment of patient prognosis.

Classification trees use binary recursive partitioning methods to partition the sample into distinct subsets [1–4]. Although their use is popular in clinical research, concerns have been raised about the accuracy of tree-based methods of classification and regression [2,4]. In the data-mining and machine-learning literature, alternatives to and extensions of classical classification trees have been developed in recent years. Many of these methods involve aggregating classifications over an ensemble of classification trees. For this reason, many of these methods are referred to as ensemble methods. Ensemble-based methods include bagged classification trees, random forests, and boosted trees. Alternate classification methods include support vector machines (SVMs).

In patients with acute heart failure (HF), there are two distinct subtypes: HF with preserved ejection fraction (HFPEF) and HF with reduced ejection fraction (HFREF). The distinction between HFPEF and HFREF is particularly relevant in the clinical setting. Although the treatment of HFREF is based on a multitude of large randomized clinical trials, the evidence base for the treatment of HFPEF is much smaller and more focused on related comorbid conditions [5]. Although the overall prognosis appears to be similar within the two subtypes of HF, there are important differences in cause-specific mortality, which would be relevant in risk stratification and disease management [6]. The diagnosis of HFREF versus HFPEF is ideally made using results from echocardiography. Although echocardiography should ideally be done in all HF patients at some point in their clinical care, this test is not always performed even in high-resource regions, and treatment decisions may need to be made before echocardiographic data are available. In one US Medicare cohort, more than one-third of HF patients did not undergo echocardiography in hospital [7].

The present study had two objectives. First, to compare the accuracy of different methods for classifying HF patients according to two disease subtypes, HFPEF vs. HFREF, and for predicting the probability of patients having HFPEF in a population-based sample of HF patients in Ontario, Canada. Second, to compare the accuracy of the prediction of the presence of HFPEF using methods from the data-mining literature with that of conventional logistic regression.

2. Methods for classification and prediction

In this section, we describe the different methods that will be used for classification and prediction. For classification, we restrict our attention to binary classification in which subjects are classified as belonging to one of two possible categories. Our case study will consist of patients with acute HF that is further classified as HF with preserved ejection fraction (HFPEF) and HF with reduced ejection fraction (HFREF). By prediction, we mean prediction of the probability of an event or of being in a particular state. In our case study, this will be the predicted probability of having HFPEF. We consider the following classification methods: classification trees, bagged classification trees, random forests, boosted classification trees, and SVMs. For prediction, we consider the following methods: logistic regression, regression trees, bagged regression trees, random forests, and boosted regression trees.